CHAPTER

2

Describing Data

Contemporary German artist Andreas Gursky (b 1955) digitally manipulated this photograph so objects in the background appear as clearly as objects in the foreground. The resulting work of art, *99 Cent* (1999), conveys the vast amount of information that surrounds you in a supermarket. What data about the inventory of this store is helpful to you as a customer? What data is helpful to the management of the store? How would you describe and summarize that data with graphs or numbers?

OBJECTIVES

In this chapter you will

- **create, interpret, and compare graphs of data sets**
- **calculate numerical measures that help you understand and interpret a data set**
- **make conclusions about a data set and compare it with other data sets based on graphs and numerical values**

LESSON 2.1

Measures of Central Tendency and Box Plots

That is what learning is. You suddenly understand something you've understood all your life, but in a new way.

DORIS LESSING

Newspapers, magazines, the evening news, commercials, government bulletins, and sports publications bombard you daily with data and statistics. As an informed citizen, you need to be able to interpret this information in order to make intelligent decisions.

In this chapter you will graph data sets in several different ways. You'll also study some numerical measures that help you better understand what a data set tells you. Each numerical measure can be called a **statistic.** A collection of measures, or the mathematical study of data collection and analysis, is called **statistics.** Studying statistics helps you learn how to collect, organize, analyze, and interpret data.

EXAMPLE A

Owen is a member of the student council and wants to present data about backpack safety to the school board. He collects these data on the weights of backpacks of 30 randomly chosen students. How much does the typical backpack weigh at Owen's school?

Student	Grade	Weight of backpack (lb)
1	Junior	10
2	Senior	20
3	Junior	9
4	Junior	17
5	Junior	3
6	Junior	10
7	Senior	15
8	Junior	15
9	Senior	7
10	Senior	10
11	Junior	9
12	Senior	10
13	Senior	9
14	Junior	7
15	Senior	4
16	Senior	6
17	Senior	7
18	Senior	9
19	Junior	13
20	Junior	10
21	Senior	8
22	Senior	7
23	Senior	4
24	Senior	4
25	Junior	8
26	Junior	33
27	Senior	10
28	Senior	9
29	Senior	7
30	Junior	16

▶ Solution

There are three statistics you could use to describe a typical item from a list of numerical data: the mean, the median, or the mode.

The **mean** is 10.2 lb, the sum of all data values, 306 lb, divided by the number of values, 30.

The **median** is 9 lb, the middle value when the data are arranged in order.

3, 4, 4, 4, 6, 7, 7, 7, 7, 7, 8, 8, 9, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10, 13, 15, 15, 16, 17, 20, 33

$$\frac{9+9}{2} = 9$$

Because there is an even number of values, the median is the mean of the two middle values.

The **mode** is 10 lb, the weight that occurs most frequently.

[▶ See **Calculator Note 1G** to learn how to enter these data into your calculator. See **Calculator Note 2B** to calculate the mean and median.◀]

You can justify using any of these three statistics as a typical weight. If Owen wants to present a statistic that implies backpacks are too heavy, he might want to use the mean because it is higher than the median or mode due to one very large data value. When a data set has one or more values that are far from the rest, the median often is more representative of the data than the mean.

Health CONNECTION

The weight of a backpack, if carried improperly, can cause physical injuries. The American Chiropractic Association recommends that the items inside a backpack weigh no more than 10% of the person's body weight. To prevent muscle aches, fatigue, and pain in the shoulders, neck, and back, use straps on both shoulders, and adjust the straps so that the backpack falls below the shoulders and rests on the hips.

The data set in Example A did not include every student in the school, so it may or may not tell much about all student backpack weights. If Owen took his sample from the first 30 students who arrived to a single class, then the data set might be biased, or unfair: it could represent students who hurry to class because their backpacks are too heavy. How might the information be biased if Owen took the sample from the first 30 volunteers?

If you assume Owen's data are from a **random sample** of *all* students, then you can make some general conclusions about all backpacks at his school. The three values commonly used to describe a typical data value-the mean, the median, and the mode-are called **measures of central tendency.**

In statistics, the mean is often referred to by the symbol $\bar{x}$ (pronounced "*x* bar"). Another symbol, $\sum$ (capital *sigma*), is used to indicate the sum of the data values.

For example, $\sum_{i=1}^{5} x_i$ means $x_1 + x_2 + x_3 + x_4 + x_5$, where x_1, x_2, x_3, x_4, and x_5 are the individual data values. So the mean of *n* data values is given by

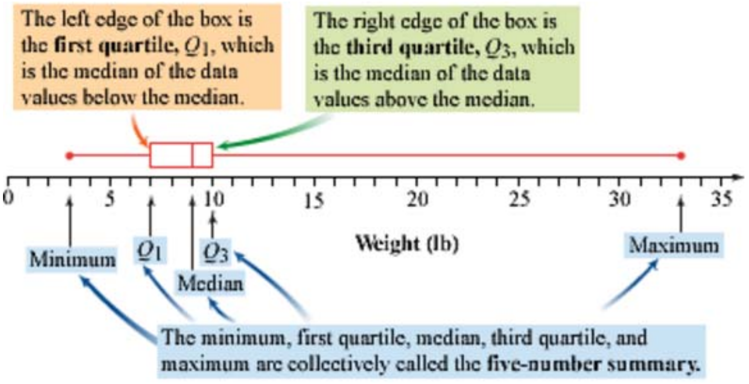
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The sum of all data values.

The number of data values.

Summarizing a data set with a single "typical" number or statistic is an incomplete picture. Sharing the entire data set is not usually informative either. A good description of the data set includes not only a measure of central tendency but the spread and distribution of the data as well. This is often done with a set of summary values or a graph.

The **box plot** (or **box-and-whisker plot**) provides a visual tool for analyzing information about a data set. This is the box plot of the backpack data from Example A. [► See Calculator Note 2C to learn how to create a box plot on your calculator.◄]


The left edge of the box is the **first quartile**, Q_1, which is the median of the data values below the median.

The right edge of the box is the **third quartile**, Q_3, which is the median of the data values above the median.

The minimum, first quartile, median, third quartile, and maximum are collectively called the **five-number summary**.

The lines extending from the "box" are called "whiskers." They identify the minimum and maximum values of the data. The difference between the maximum and minimum is the **range** of the data.

EXAMPLE B

Use the box plot above to analyze the backpack data.

a. What percentage of the data values is represented by the lower whisker?

b. What are the values for the first quartile, the median, and the third quartile?

c. What is the five-number summary for this data set?

► Solution

Read the data values from the box plot at the five-number summary points, and use the definition of quartile.

a. One-quarter, or 25%, of the data values are represented by the lower whisker. As a matter of fact, one-quarter of the data values are represented by the upper whisker, one-quarter are represented by the upper part of the box, and one-quarter are represented by the lower part of the box as well.

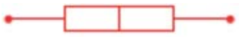
b. There are 30 values, so after arranging them in order, the first quartile is the eighth value, or 7 lb; the median is the mean of the fifteenth and sixteenth values, or 9 lb; and the third quartile is the twenty-third value, or 10 lb.

c. The five-number summary is 3, 7, 9, 10, 33.

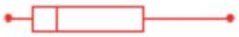
History CONNECTION

One early large-scale statistical survey was that of the 16th -century Hawaiian king, Umi. According to legend, he collected all of his people on a small plain, afterward called the Plain of Numbering, and asked each person to deposit a stone in an area encircling the temple on that plain. The stones were placed in piles according to district, and the piles were located in the direction of the districts. The result showed the relative sizes of the districts' populations.

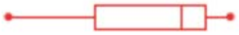
Statisticians often talk about the shape of a data set. **Shape** describes how the data are distributed relative to the center. A **symmetric** data set is balanced, or nearly so, at the center. Note that it does not have to be exactly equal on both sides to be called symmetric. **Skewed** data are spread out more on one side of the center than on the other side. The backpack data is an example of skewed data. You will learn more about spread in the next lesson. For now, a box plot can be a good indicator of shape because the median is clearly visible as the center.



This box plot shows a symmetric data set.



Skewed right implies that the data are spread more to the right of the center than to the left.



This data set is skewed left.

Investigation

Pulse Rates

You will need

- a watch or clock with a second hand

Pulse rate is often used as a measure of whether or not a person is in good physical condition. In this investigation you will practice making box plots, compare box plots, and draw some conclusions about pulse rates.

Step 1 Measure and record your resting pulse for 15 s. Multiply this value by 4 to get the number of beats per minute. Pool data from the entire class.

Step 2 Exercise for 2 min by doing jumping jacks or by running in place. Afterward, measure and record your exercise pulse rate. Pool your data.

Step 3 Order each set of data. Calculate the five-number summaries for your class's resting pulse rates and for your exercise pulse rates.

Step 4 Prepare a box plot of the resting pulse rates and a box plot of the exercise pulse rates. Determine a range suitable for displaying both of these graphs on a single axis.

Step 5 Draw conclusions about pulse rates by comparing these two graphs. Be sure to compare not only centers but also spreads and shapes. How could a physician or a personal trainer use your results to determine whether a client is in good physical condition?

Box plots are a convenient way to compare two data sets. Not only can you readily compare the medians, but you can also see if the two sets are distributed in the same manner.

EXERCISES

Practice Your Skills

1. Find the mean, median, and mode for each data set.

a. Time for pizza delivery (min): {28, 31, 26, 35, 26}

b. Yearly rainfall (cm): {11.5, 17.4, 20.3, 18.5, 17.4, 19.0}

c. Cost of a small popcorn at movie theaters ($): {2.75, 3.00, 2.50, 1.50, 1.75, 2.00, 2.25, 3.25}

d. Number of pets per household: {3, 2, 1, 0, 3, 4, 1}

2. A data set has a mean of 12 days, the median is 14 days, and there are three values in the data set.

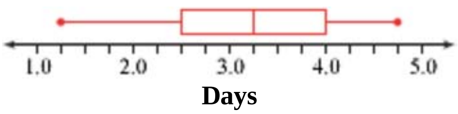
a. What is the sum of all three data values?

b. What is the one value you know?

c. Create a data set that has the statistics given. Is there more than one data set that could have these statistics?

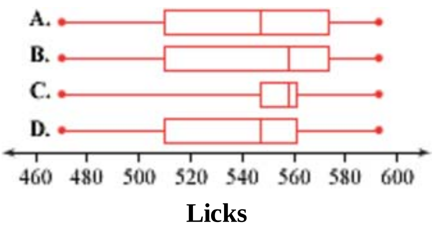
3. Approximate the values of the five-number summary for this box plot. Give the full name for each value.

Life Span of House Flies



4. Match this data set to one of the four box plots. Licks to the center of a lollipop: {470, 510, 547, 558, 561, 574, 593}

Licks to the Center of a Lollipop



5. Match this box plot to one of the data sets for the number of minutes on the phone spent by 13 customer service representatives in a given hour.

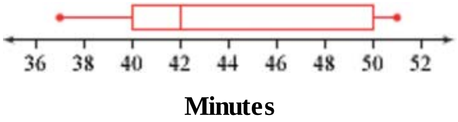
A. {36, 37, 38, 39, 40, 41, 42, 43, 44, 46, 48, 50, 52}

B. {37, 40, 42, 50, 51, 51, 51, 51, 51, 51, 51, 51, 51}

C. {36, 37, 40, 40, 40, 42, 42, 43, 44, 50, 50, 51, 52}

D. {37, 39, 40, 40, 40, 41, 42, 43, 44, 49, 51, 51, 51}

Time on the Phone



Keymath.com More Practice Your Skills

Reason and Apply

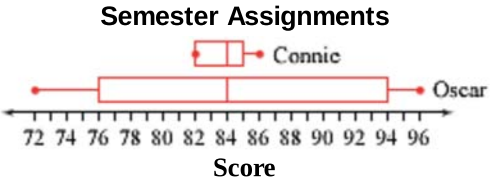
6. Here are the scores on semester assignments for two students.

Connie: {82, 86, 82, 84, 85, 84, 85}

Oscar: {72, 94, 76, 96, 90, 76, 84}

Find the mean and median for each set of scores, and explain why they do not tell the whole story about the differences between Connie's and Oscar's scores.

7. These box plots represent Connie's and Oscar's scores from Exercise 6.



Write a paragraph describing the information pictured in the box plots. Use the box plots to help you draw some statistical conclusions. In your description, include answers to such questions as What does it mean that the second box plot is longer? Where is the left whisker of the top box plot? What does it mean when the median isn't in the middle of the box? What does it mean when the left whisker is longer than the right whisker?

8. Homer Mueller has played in the minor leagues for 11 years. His home run totals, in order, for those years are 56, 62, 49, 65, 58, 52, 68, 72, 25, 51, and 64.

a. Construct a box plot showing Homer's data.

b. Give the five-number summary.

c. Find $\bar{x}$.

d. How many home runs would Homer need to hit next season to have a 12-year mean of 60?

Pedro Martínez pitches for the Boston Red Sox.

Sports CONNECTION

Baseball fans thrive on the "stats" of their favorite players. For example, earned run average (ERA) is a statistic calculated for pitchers. You calculate ERA by taking the number of earned runs scored on the pitcher and dividing it by one-ninth the total number of innings pitched.

9. The **interquartile range** ***(IQR)*** is the difference between the first and third quartiles, or the length of the box in a box plot.

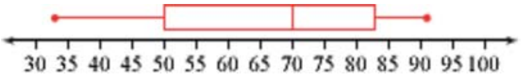
a. Look at the box plots in Exercise 7. What are the range and interquartile range for Connie? For Oscar?

b. Find the range and interquartile range for your box plot from Exercise 8.

10. Invent a data set with seven distinct values and a mean of 12.

11. Invent a data set with seven values and a mode of 70 and a median of 65.

12. Invent a data set with seven values that creates this box plot.



13. Refer to the backpack data listed in Example A. Separate the data by grade level.

a. Compute the mean weight for the juniors and for the seniors.

b. Calculate the median for each grade level.

c. Compare the mean and median values for each grade level. Which is the greater value, mean or median, in each set?

14. Use the backpack data separated by grade level from Exercise 13.

a. Create a box plot for each grade level. Put both box plots on the same axis.

b. Based on the information in your box plots, write a brief statement analyzing these two groups. Use the vocabulary developed in this lesson in your statement.

c. Based on your box plots, explain why the means or medians may have been greater in 13c.

15. **APPLICATION** Lord Rayleigh was one of the early pioneers in studying the density of nitrogen. (Read the Science Connection below.) The following are data that he collected. Lord Rayleigh's measurements first appeared in *Proceedings of the Royal Society of London* (London, vol. 55, 1894). Each piece of data is the mass in grams of nitrogen filling a certain flask under a specified temperature and pressure.

Mass of Nitrogen Produced from Chemical Compounds (g)		
2.30143	2.29890	2.29816
2.30182	2.29869	2.29940
2.29849	2.29889	2.30074
2.30054		

Mass of Nitrogen Produced from the Atmosphere (g)		
2.31017	2.30986	2.31010
2.31001	2.31024	2.31010
2.31028	2.31163	2.30956

a. Calculate the five-number summary for each set of data.

b. On the same axis, create a box plot for each set of data.

c. Describe any similarities and differences in the shapes of the box plots. Do the box plots support Lord Rayleigh's conjecture?

Science CONNECTION

One of the earliest persons to study the density of nitrogen was the English scientist Lord Rayleigh (1842-1919, born John William Strutt). Working with fairly small samples, he noticed that the density of nitrogen produced from chemical compounds was different from the density of nitrogen produced from the atmosphere. On the supposition that the air-derived gas was heavier than the "chemical" nitrogen, he conjectured the existence in the atmosphere of an unknown ingredient. In 1894, Lord Rayleigh isolated the unknown ingredient, the colorless, tasteless, and odorless gas called argon. In 1904, Lord Rayleigh was awarded the Nobel Prize in physics for his discovery. You can learn more about Rayleigh's work by using the links at www.keymath.com/DAA .

Lord Rayleigh

Review

16. Find the next three terms in each sequence and write a recursive formula.

a. 42, 45, 48 . . .

b. 16, 40, 100, . . .

17. Evaluate each expression. Write your answers both in radical form and in decimal form rounded to one decimal place.

a. $\sqrt{\frac{432}{6}}$

b. $\sqrt{\frac{782+1354}{24}}$

c. $\sqrt{\frac{49+121+16+81+100}{4}}$

18. **APPLICATION** Rebecca wants to buy a used drum set that costs $400. Either she can buy it now on credit and pay an annual interest rate of 15%, compounded monthly, on the unpaid balance, or she can wait until she has saved the money in her bank account which earns an annual interest rate of 5% compounded monthly. Either way she can contribute $40 each month from her weekend job. How long will it take to pay for the drum set if she buys it on credit? How long would she have to save to buy it? Give Rebecca advice.

19. Solve.

a. $x + 8 = 15$

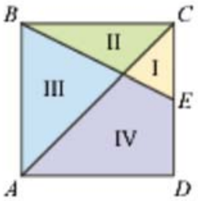
b. $3x = 15$

c. $3x + 8 = 15$

IMPROVING YOUR GEOMETRY SKILLS

Dissecting a Square

Square $ABCD$ is divided into four regions by drawing $\overline{AC}$ and $\overline{BE}$, where E is the midpoint of $\overline{CD}$. Compare the areas of regions I, II, III, and IV.



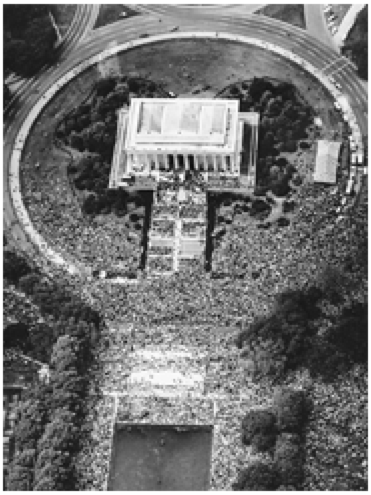
LESSON 2.2

Measures of Spread

Out on the edge you see all kinds of things you can't see from the center.

KURT VONNEGUT

If you ask several people to estimate the number of people in a crowd, their estimates will usually differ. The mean or median would measure the central tendency of the estimates, but neither of these statistics tell how widely people's estimates differ. Measuring variability, or **spread,** in numerical data allows a more complete description than just stating a measure of central tendency. In this lesson you will investigate different ways to measure and describe variability.



On August 28, 1963, about 250,000 people gathered at the Lincoln Memorial to support civil rights legislation. It was here that Martin Luther King, Jr., gave his "I Have a Dream" speech. The March on Washington was the largest gathering of people anywhere to that date.

Investigation

A Good Design

You will need

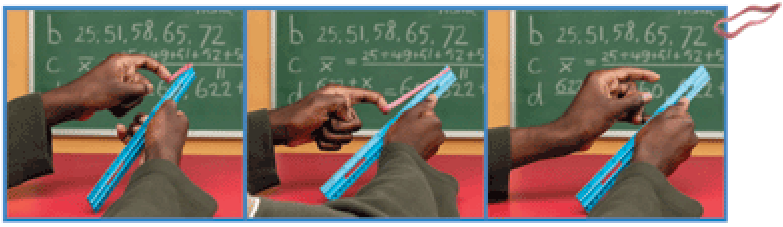
- a rubber band
- a ruler
- a measuring tape or metersticks
- paper
- books
- a pad of paper or cardboard

In a well-designed experiment, you should be able to follow a specific procedure and get very similar results every time you perform the experiment. In this investigation, you will attempt to control the setup of an experiment in order to limit the variability of your results.

Select and perform one of these experiments. Make complete and careful notes about the setup of your experiment.

Experiment 1: Rubber Band Launch

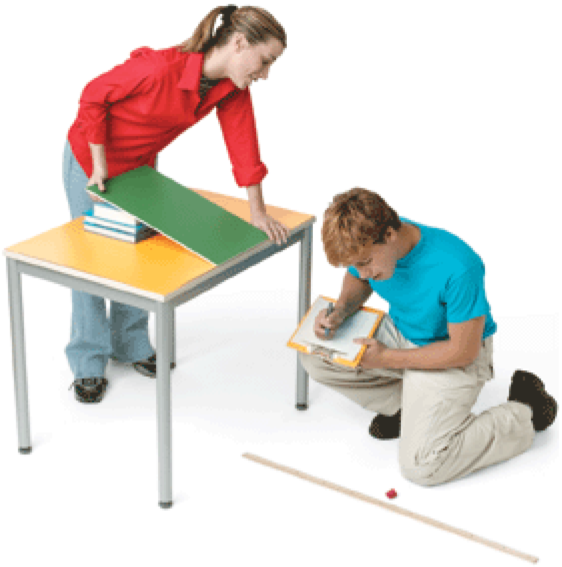
In this experiment you'll use a ruler to launch a rubber band. Select the height and angle of your launch and the length of your stretch, and determine any other factors that might affect your results. Launch the rubber band into an area clear of obstructions. Record the horizontal distance of the flight. Repeat this procedure as precisely as you can with the same rubber band, the same launch setup, and the same stretch another seven or eight times.



© 2004 Key Curriculum Press

Experiment 2: Rolling Ball

In this experiment you'll roll a ball of paper down a ramp and off the edge of your desk. Build your ramp from books, notebooks, or a pad of paper. Select the height and slope of your ramp and the distance from the edge of your desk, and determine any other factors that might affect your results. Make a ball by crumpling a piece of paper, and roll it down the ramp. Record the horizontal distance to the place where the ball hits the floor. Repeat this procedure with the same ball, the same ramp setup, and the same release another seven or eight times.



Step 1 Use your data from Experiment 1 or 2.
Calculate the mean distance for your trials.

Step 2 On average, how much do your data values differ from the mean? How does the variability in your results relate to how controlled your setup was? Determine a way to calculate a *single* value that tells how accurate your group was at repeating the procedure. Write a formula to calculate your statistic.

Step 3 There is a value known as the **standard deviation** that helps measure the spread of data away from the mean. Use your calculator to find this value for your data. [► See Calculator Note 2B to learn how to calculate the standard deviation using your calculator.◄]

Step 4 See if you can find a formula or procedure that allows you to calculate the value of the standard deviation by hand. (Hints: A **deviation** is the difference between a data value and the mean of the data set. Standard deviation involves both squaring and taking a square root.)

Step 5 Repeat the experiment and collect another set of data from seven or eight trials. Calculate your statistic and the standard deviation for your new set of data. How do the results of your experiments compare? Write a report explaining your procedures and conclusions. In your report, explain some things that you could do differently in order to minimize the standard deviation.

If you want to measure the spread of data, it is typical to start by finding the mean. Next you find the **deviation,** or directed distance, from each data value to the mean. When you compare the results from two groups in the previous investigation, you may find that one group's mean is 200 cm and another group's mean is 300 cm. However, if their deviations are similar, then they performed the experiment equally well. The deviations let you compare the spread independent of the mean.

Consider two groups that each do the rubber band launch seven times.

Group A distance (cm): {182, 186, 182, 184, 185, 184, 185}

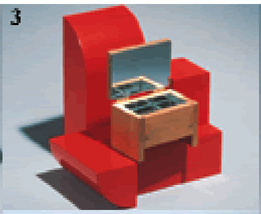
Group B distance (cm): {152, 194, 166, 216, 200, 176, 184}

The mean for each group is 184. The individual deviations, $x_i - \bar{x}$, for each data value, x_i , are in the table below.

Group A			Group B		
Data value	**Distance**	**Deviation**	**Data value**	**Distance**	**Deviation**
x_1	182	182 – 184 = **–2**	x_1	152	152 – 184 = **–32**
x_2	186	186 – 184 = **2**	x_2	194	194 – 184 = **10**
x_3	182	182 – 184 = **–2**	x_3	166	166 – 184 = **–18**
x_4	184	184 – 184 = **0**	x_4	216	216 – 184 = **32**
x_5	185	185 – 184 = **1**	x_5	200	200 – 184 = **16**
x_6	184	184 – 184 = **0**	x_6	176	176 – 184 = **– 8**
x_7	185	185 – 184 = **1**	x_7	184	184 – 184 = **0**

Begun in 1992 by Lorraine Serena and Elena Siff, *Women Beyond Borders* is a worldwide art project in which women artists are given identical wooden boxes and asked to transform them. If you consider the original box to be the mean, some of the resulting artworks have large deviations.

These deviations show more variation in Group B's distances than in Group A's distances. That might imply Group B's experiment was not designed well enough to give consistent results.



From left: **1.** The original pine box; **2.** Darlene Nguyen-Ely, USA, Vietnam, *Journey #17;* **3.** Madoka Hirata, Japan, *The Distance from Time #1;* **4.** Elena Mary Siff, USA, *Narcissism;* **5.** Cirenaica Moreira Diaz, Cuba, *Untitled;* **6.** Alejandra Mastro Sesenna, Guatemala, *Eva's Last Wish;* **7.** Gordana Kaljalovic Odanovic, Yugoslavia, *Model of Intimacy;* **8.** Lilian Nabulime, Kenya, *My Self.* On the web at *www.womenbeyondborders.org.*

How can you combine the deviations into a single value that reflects the spread in a data set? Finding the sum is a natural choice. However, if you think of the mean as a balance point in a data set, then the directed distances above and below the mean should balance out. Hence, the deviation sum for both Group A and Group B is zero.

Group A's deviation sum: –2 + 2 + –2 + 0 + 1 + 0 + 1 = 0
Group B's deviation sum: –32 + 10 + –18 + 32 + 16 + –8 + 0 = 0

In order for the sum of the deviations to be useful, you need to eliminate the effect of the different signs. Squaring each deviation is one way to do this.

	Group A			Group B		
	Distance	Deviation	(Deviation)^2	Distance	Deviation	(Deviation)^2
	182	-2	**4**	152	–32	**1024**
	186	2	**4**	194	10	**100**
	182	-2	**4**	166	–18	**324**
	184	0	**0**	216	32	**1024**
	185	1	**1**	200	16	**256**
	184	0	**0**	176	– 8	**64**
	185	1	**1**	184	0	**0**
	sum = **14**			*sum* = **2792**		
Variance	$\frac{sum}{6} = \mathbf{2.\overline{3}}$			$\frac{sum}{6} = \mathbf{465.\overline{3}}$		
Standard deviation	$\sqrt{\frac{sum}{6}} \approx \mathbf{1.5}$			$\sqrt{\frac{sum}{6}} \approx \mathbf{21.6}$		

When you sum the squares of the deviations, the sum is no longer zero. The sum of the squares of the deviations, divided by one less than the number of values, is called the **variance** of the data. The square root of the variance is called the **standard deviation** of the data. The standard deviation provides one way to judge the "average difference" between data values and the mean. It is a measure of how the data are spread around the *mean*.

Standard Deviation

The **standard deviation,** *s*, is a measure of the spread of a data set.

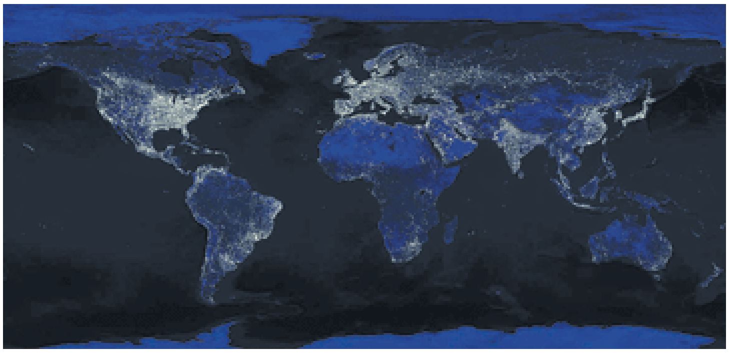
$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

where x_i represents the individual data values, *n* is the number of values, and $\bar{x}$ is the mean. The standard deviation has the same units as the data.

The larger standard deviation for Group B indicates that their distances generally lie much farther from the mean than do Group A's. A large value for the standard deviation tells you that the data values are not as tightly packed around the mean. As a general rule, a set with more data near the mean will have less spread and a smaller standard deviation.

You may wonder why you divide by $(n - 1)$ when calculating standard deviation. As you know, the sum of the deviations is zero. So, if you know all but one of the deviations, you can calculate the last deviation by making sure the sum will be zero. The last deviation depends on the rest, so the set of deviations contains only $(n - 1)$ independent pieces of data.

When you make a box plot, you have a visual representation of how data are spread around the *median*. The range (the distance between the whisker endpoints) and the interquartile range (the length of the box) are measures of spread around the median. **Outliers** are data values that differ significantly from the majority of the data. The exact definition of an outlier may vary according to different textbooks. Some statisticians identify outliers from a box plot as values that are more than $1.5 \cdot IQR$ from either end of the box.



The National Geophysical Data Center created this map of human settlements, based on composite data from many satellites about sources of nighttime light. You could consider isolated points of light to be outliers because they are far removed from dense concentrations of light.

EXAMPLE

This table gives the student-to-teacher ratios for public elementary and secondary schools in the United States.

Student-to-Teacher Ratios for Public Elementary and Secondary Schools (2000-2001 School Year)

State	AK	AL	AR	AZ	CA	CO	CT	DE	FL	GA	HI	IA	ID
Ratio	16.9	15.4	14.1	19.8	20.6	17.3	13.7	15.3	18.4	15.9	16.9	14.3	17.9

State	IL	IN	KS	KY	LA	MA	MD	ME	MI	MN	MO	MS	MT
Ratio	16.1	16.7	14.4	16.8	14.9	14.5	16.3	12.5	18.0	16.0	14.1	16.1	14.9

State	NC	ND	NE	NH	NJ	NM	NV	NY	OH	OK	OR	PA	RI
Ratio	15.5	13.4	13.6	14.5	13.1	15.2	18.6	13.9	15.5	15.1	19.4	15.5	14.8

State	SC	SD	TN	TX	UT	VA	VT	WA	WI	WV	WY
Ratio	14.9	13.7	14.9	14.8	21.9	12.5	12.1	19.7	14.1	13.7	13.3

(U.S. Department of Education, National Center for Education Statistics)

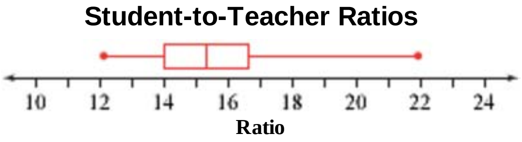
a. Calculate the mean and the standard deviation. What do the statistics tell you about the spread of the student-to-teacher ratios?

b. Make a box plot of the data and identify any outliers.

Solution

a. The mean student-to-teacher ratio is approximately 15.63. The standard deviation is approximately 2.18. So, many of the data values are within 2.18 units of the mean of 15.63.

b. The five-number summary is 12.1, 14.1, 15.15, 16.8, 21.9.



The interquartile range is 2.7. To be an outlier, a data value must be 1.5 · 2.7, or 4.05, away from an end of the box. That means it must be less than 14.1 – 4.05, or 10.05, or more than 16.8 + 4.05, or 20.85. There is one ratio, 21.9 (UT), that meets this condition, so it is an outlier. [▸ Revisit Calculator Note 2C to learn how to make a box plot that shows outliers. ◂]

As you work the exercises, you may notice that finding a measure of central tendency is usually an integral step of measuring the spread. In order to calculate the standard deviation you first need to calculate the mean. The interquartile range relies on the first and third quartiles, which in turn rely on the median.

EXERCISES

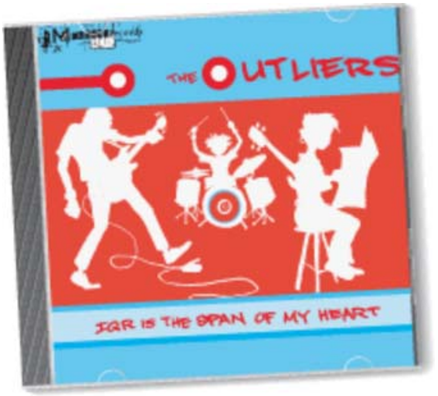
Practice Your Skills

1. Given the data set {41, 55, 48, 44}:

a. Find the mean.

b. Find the deviation from the mean for each value.

c. Find the standard deviation of the data set.



2. The lengths in minutes of nine music CDs are 45, 63, 74, 69, 72, 53, 72, 73, and 50.

a. Find the mean.

b. Find the deviation from the mean for each value.

c. Find the standard deviation of the data set.

d. What are the units of the mean, the deviations from the mean, and the standard deviation?

3. In a classroom experiment, 11 bean plants were grown from seeds. After two weeks, the heights in centimeters of the plants were 9, 10, 10, 13, 13, 14, 15, 16, 17, 19, and 21.

a. Find the five-number summary.

b. Find the range and interquartile range.

c. What are the units of the range and interquartile range?

4. In order to monitor weight, a cookie manufacturer samples jumbo chocolate chip cookies as they come off the production line. The weights in grams of 11 cookies are 22, 30, 27, 35, 32, 28, 18, 22, 25, 30, and 28.

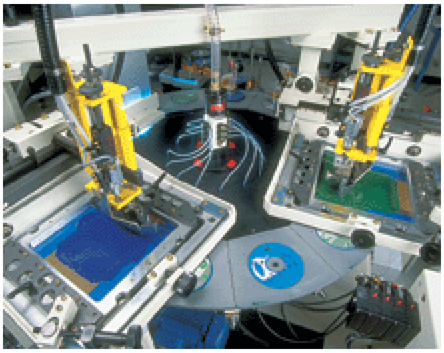
a. Find the five-number summary.
b. Find the range and interquartile range.
c. What are the units of the range and interquartile range?

Keymath.com More Practice Your Skills

5. Invent a data set with seven data values such that the mean and the median are both 84, the range is 23, and the interquartile range is 12.

Reason and Apply

6. **APPLICATION** The mean diameter of a Purdy Goode Compact Disc is 12.0 cm, with a standard deviation of 0.012 cm. No CDs can be shipped that are more than one standard deviation from the mean. How would the company's quality control engineer use those statistics?



This automated machine paints labels on compact discs.

7. Some statisticians identify outliers as data values that are more than two standard deviations, or 2*s*, from the mean. Use this method to identify any outliers in the student-to-teacher ratios from the example on page 89. How do the outliers found by standard deviation compare to the outliers found by interquartile range?

8. Find the standard deviation and interquartile range of the backpack data from Example A in Lesson 2.1. Which of these two values is larger? Will this value always be larger? Explain your reasoning and find or create another data set that supports your answer.

9. Two data sets have the same range and interquartile range but the first is symmetric and the second is skewed left.

a. Sketch two box plots that satisfy the conditions for the two sets.
b. Would you guess that the standard deviation of the skewed data set is less than, more than, or the same as the first? Explain your reasoning.
c. Invent two data sets of seven values each that satisfy the conditions.
d. Find the standard deviation for the two data sets. Do the standard deviations support your answer to 9b?

10. Students collected eight length measurements during a mathematics lab. The mean measurement was 46.3 cm, and the deviations of *seven* individual measurements were 0.8 cm, – 0.4 cm, 1.6 cm, 1.1 cm, –1.2 cm, –0.3 cm, and –1.0 cm.

a. What were the original eight measurements collected?
b. Find the standard deviation of the original measurements.
c. Which measurements are more than one standard deviation above or below the mean?

11. **APPLICATION** The students in four classes recorded their resting pulse rates in beats per minute. The class means and standard deviations are given at right.

a. Which class has students with pulse rates most alike? How can you tell?

b. Can you tell which class has the students with the fastest pulse rates? Why or why not?

Resting Pulse Rates (beats/min)

Class	Mean	Standard deviation
First period	79.4	3.2
Third period	74.6	5.6
Fifth period	78.2	4.1
Sixth period	80.2	7.6

12. Here are the mean daily temperatures in degrees Fahrenheit for two cities.

a. Find the mean and standard deviation for each city.

b. Draw a box plot for each city. Find each median and interquartile range.

c. Which city has the smaller spread of temperatures? Justify your conclusion.

d. Does the interquartile range or the standard deviation give a better measure of the spread? Justify your conclusion.

Mean Temperatures (°F)

Month	Juneau, Alaska	New York, New York
January	24	31
February	28	33
March	33	41
April	40	51
May	47	60
June	53	69
July	56	76
August	55	75
September	49	68
October	42	57
November	32	47
December	27	37

13. Members of the school mathematics club sold packages of hot chocolate mix to raise funds for their club activities. The numbers of packages sold by individual members are given at right.

a. Find the median and interquartile range for this data set.

b. Find the mean and standard deviation.

c. Draw a box plot for this data set. Use the $1.5 \cdot IQR$ definition to name any numbers that are outliers.

d. Remove the outliers from the data set and draw another box plot.

e. With the outliers removed, recalculate the median and interquartile range and the mean and the standard deviation.

f. Which is more affected by outliers, the mean or the median? The standard deviation or the interquartile range? Explain why you think this is so.

65	76	100	67	44
147	82	94	92	79
158	77	62	85	71
69	88	80	63	75
62	68	71	73	74

14. Refer to the data in Exercise 13.

a. Suppose each package of hot chocolate yields a net profit of 28¢. Draw a box plot for the profit each club member generates. Find the median and interquartile range, and the mean and standard deviation for the profits. Compare your profit statistics with the original statistics describing the numbers of packages sold. How are the two sets of statistics and the graphs related?

b. Use your findings from 14a to predict the net profit statistics if the net profit per package is 35¢.

c. Suppose the school audit finds that each individual member actually sold 20 packages fewer than originally reported. Find the median and interquartile range, and the mean and standard deviation. Describe a process you could use to find the corrected results for all of the information requested in Exercise 13.

d. Use your findings from 14c to predict the statistical results if instead there were 10 packages fewer per member than originally reported.

15. **APPLICATION** Matt Decovsky wants to buy a 160W CD player for his car at an online auction site. Before bidding, he decides to do some research on the selling price of recently sold CD players. His search comes up with these 20 prices.

$74.00	$102.50	$64.57	$74.00
$82.87	$73.01	$77.00	$71.00
$71.01	$112.50	$86.00	$102.50
$76.00	$56.00	$135.50	$66.00
$71.00	$76.00	$51.00	$88.00

a. Find the mean, median, and mode.
b. Draw a box plot of the data. Describe its shape.
c. Find the interquartile range and determine whether there are any outliers.
d. While looking over the items' descriptions, Matt realizes that the outlier CD players contain features that don't interest him. If he removes the outliers, will the mean or the median be less affected? Explain.
e. Sketch a new box plot with the outliers removed. How does this help you support your answer in 15d?
f. If you were Matt, what might you set as a target price in your bidding? Explain your reasoning.

Review

16. Celia lives 2.4 km from school. She misses the school bus and starts walking at 1.3 m/s. She has 20 min before school starts. Write a recursive formula and use it to find out whether or not she gets to school on time.

17. Solve.

a. $\frac{x+5}{4} + 3 = 19$

b. $\frac{3(y-4)+6}{6} - 2 = 7$

18. These data sets give the weights in pounds of the offensive and defensive teams of the 2002 Super Bowl Champion New England Patriots. *(www.nfl.com)*

Offensive players' weights (lb): {190, 305, 310, 320, 315, 322, 255, 190, 220, 245, 230}
Defensive players' weights (lb): {280, 305, 280, 270, 250, 253, 245, 199, 196, 207, 218}

a. Find the mean and median weights of each team.
b. Prepare a box plot of each data set. Use the box plots to make general observations about the differences between the two teams.

LESSON 2.3

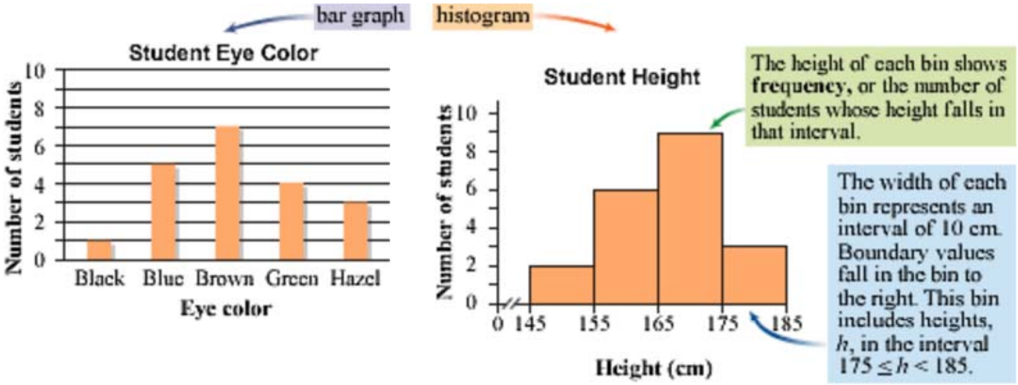
Histograms and Percentile Ranks

You miss 100 percent of the shots you never take.

WAYNE GRETZKY

A box plot gives you an idea of the overall distribution of a data set, but in some cases you might want to see other information and details that a box plot doesn't show. A **histogram** is a graphical representation of a data set, with columns to show how the data are distributed across different intervals of values. Histograms give vivid pictures of distribution features, such as clusters of values, or gaps in data.

The columns of a histogram are called **bins** and should not be confused with the bars of a bar graph. The bars of a bar graph indicate categories-how many data items either have the same value or share a characteristic. The bins of a histogram indicate how many numerical data values fall within a certain interval. You would use a bar graph to show how many people in your class have various eye colors, but a histogram to show how many people's heights fall within various intervals.

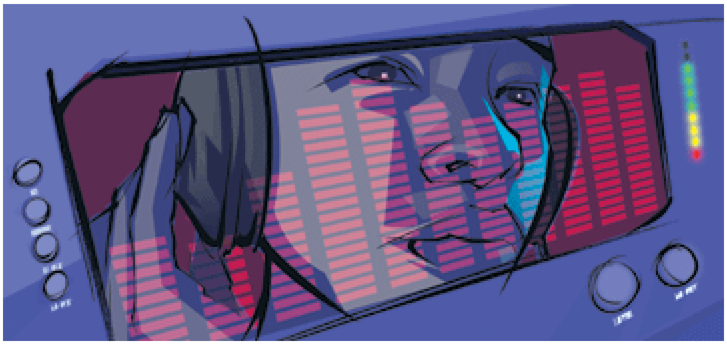


The height of each bin shows **frequency**, or the number of students whose height falls in that interval.

The width of each bin represents an interval of 10 cm. Boundary values fall in the bin to the right. This bin includes heights, h, in the interval $175 \leq h < 185$.

Histograms are a good way to display information from large data sets. Although you can't see individual data values, you can see the shape of the data and how the values are distributed throughout the range. As you will see in Example A, bin width depends on how much detail you want to show, but all the bins should have the same width.

Some stereo equalizers have spectrum displays that resemble histograms. These displays are similar to histograms because they show the output frequencies by intervals, or bands. They are different because the bands may not represent equal intervals.



EXAMPLE A

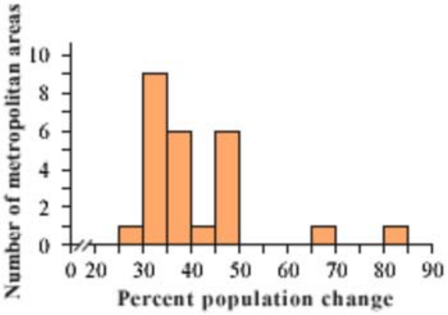
Both Graph A and Graph B were constructed from the data set in the table.

The 25 Fastest-Growing Metropolitan Areas in the United States (1990-2000)

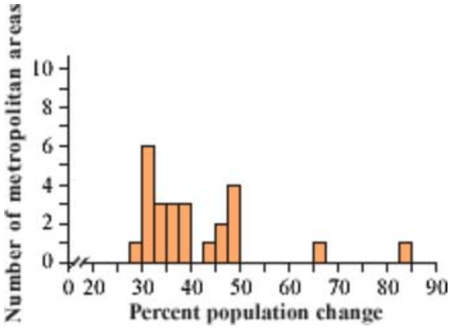
Metropolitan area	1990 Population	2000 Population	Percent Population change
Las Vegas, NV	852,737	1,563,282	83.3
Naples, FL	152,099	251,377	65.3
Yuma, AZ	106,895	160,026	49.7
McAllen, TX	383,545	569,463	48.5
Austin, TX	846,227	1,249,763	47.7
Fayetteville, AR	210,908	311,121	47.5
Boise City, ID	295,851	432,345	46.1
Phoenix, AZ	2,238,480	3,251,876	45.3
Laredo, TX	133,239	193,117	44.9
Provo, UT	263,590	368,536	39.8
Atlanta, GA	2,959,950	4,112,198	38.9
Raleigh, NC	855,545	1,187,941	38.9
Myrtle Beach, SC	144,053	196,629	36.5
Wilmington, NC	171,269	233,450	36.3
Fort Collins, CO	186,136	251,494	35.1
Orlando, FL	1,224,852	1,644,561	34.3
Reno, NV	254,667	339,486	33.3
Ocala, FL	194,833	258,916	32.9
Auburn, AL	87,146	115,092	32.1
Fort Myers, FL	335,113	440,888	31.6
West Palm Beach, FL	863,518	1,131,184	31.0
Bellingham,WA	127,780	166,814	30.5
Denver, CO	1,980,140	2,581,506	30.4
Colorado Springs, CO	397,014	516,929	30.2
Dallas, TX	4,037,282	5,221,801	29.3

(U.S. Census Bureau)

Graph A
Fastest-Growing Metropolitan Areas



Graph B
Fastest-Growing Metropolitan Areas



a. What is the range of the data?

b. What is the bin width of each graph?

c. Use the information in the table to create the same graphs on your calculator.

d. How can you know if the graph accounts for all 25 metropolitan areas?

e. Why are the columns shorter in Graph B?

f. Describe how each graph illustrates clusters and gaps in the data.

g. How many of these metropolitan areas grew between 35% and 40%? How can you tell this from each graph?

h. In what interval of Graph A is the median growth rate? If you use Graph B to answer the question, can your answer be more accurate?

i. What percentage of these metropolitan areas had population changes less than 35%?

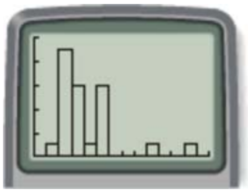
This photo of a housing development in Las Vegas conveys the city's rapid growth.

Solution

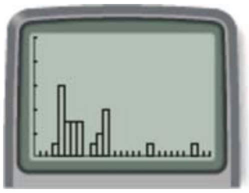
a. Population changes for these metropolitan areas range from 29.3% for Dallas, TX, to 83.3% for Las Vegas, NV. The range is 54%.

b. The width of each bin is 5% in Graph A and 2.5% in Graph B.

c. The range and the bin width are important pieces of information that you need in order to duplicate these graphs. [See Calculator Note 2D to learn how to make histograms on your calculator.]



[20, 90, 5, 0, 10, 2]



[20, 90, 2.5, 0, 10, 2]

d. Add the bin frequencies in either histogram to verify that they sum to 25.

e. The bin width of Graph B is half the bin width of Graph A. So the data values represented by each bin in Graph A get split into two bins in Graph B, making the bins generally shorter.

f. The gap between 50% and 65% in either graph means no metropolitan area had a population change between 50% and 65%. With Graph B, you also see gaps between 40% and 42.5%, and between 67.5% and 82.5%. The cluster of bins on the left side of either graph shows that most of these metropolitan areas grew between 25% and 50%.

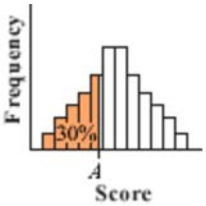
g. Six metropolitan areas grew between 35% and 40%. In Graph A, you read the frequency of the bin in the 35-to-40 interval. In Graph B, you add the frequencies of the 35-to-37.5 bin and the 37.5-to-40 bin.

h. Each graph represents 25 metropolitan areas. You find the median, which is the "middle city" or the 13th data value, by adding the frequencies of each bin from the left until you get to 13. With Graph A, the median is between 35% and 40%. With Graph B, you can narrow down the interval-the median is between 35% and 37.5%.

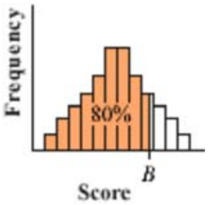
i. Add the bin frequencies to the left of 35%. You find that 10 of these 25 metropolitan areas, or 40%, had population changes less than 35%.

The **percentile rank** of a data value in a large distribution gives the percentage of data values that are below the given value. In the previous example, Fort Collins, CO, has a percentile rank of about 40 among this group, because its population change exceeds that of 40% of the metropolitan areas in this group, as you found in part i.

Suppose a large number of students take a standardized test, such as the SAT. There are so many individual scores that it would be impractical to look at all of the actual numbers. A percentile rank gives a good indication of how one person's score compares to other scores across the country.



Students with score *A* are at the 30th percentile, because their score is better than the scores of 30% of the tested students.



Likewise, students with score *B* are at the 80th percentile, because 80% of the tested students have scores that are lower.

Cultural CONNECTION

During the 15th and 16th centuries, the Inca used *quipus,* a system of knotted cords, to record numerical information, such as population. The number of knots, the number of strings, and variations in color and thickness combined to create a sophisticated method for keeping data and statistics, which scholars have yet to fully decipher.

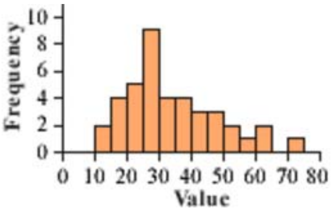
The illustration at left, from a 17th-century letter by Felipe Guáman Poma de Ayala, shows an Incan treasurer holding a *quipu.* The photo shows an actual Incan *quipu.*

EXAMPLE B

The data used in this histogram have a mean of 34.05 and a standard deviation of 14.68.

a. Approximate the percentile rank of a value two standard deviations above the mean.

b. Approximately what percentage of the data values are within one standard deviation of the mean?



Solution

Add the bin frequencies to find that there are 40 data values in all.

a. The value of two standard deviations above the mean is $34.05 + 2 \cdot 14.68$, or 63.41. All of the data values in the ten bins up to the value of 60 are less than 63.41. Adding the bin frequencies up to 60 gives 37. Therefore $\frac{37}{40}$, or approximately 92.5%, of the data lies below 63.41. So 63.41 is approximately the 93rd percentile.

b. One standard deviation above the mean is 48.73, and one standard deviation below the mean is 19.37. This interval includes at least those values in the bins from 20 to 45. So $\frac{25}{40}$, or approximately 62.5%, of the data lie within one standard deviation of the mean.

Combining what you know about measures of central tendency and spread with different displays of data enables you to provide a complete picture of a data set. The following investigation gives you an opportunity to analyze data using all of the statistics and graphs you have learned about.

Investigation

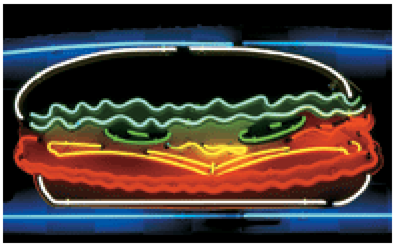
Eating on the Run

Teenagers require about 2200 to 3000 calories per day, depending on their growth rate and level of activity. The food you consume as part of your 2200- to 3000-calorie diet should include a high level of protein, moderate levels of carbohydrates and fat, and as little sodium, saturated fat, and cholesterol as possible. The table shows the minimum amount of protein and the maximum amount of other nutrients in a healthy 2500-calorie diet.

Nutrition Recommendations for a 2500-Calorie Diet

Fat	80 g
Cholesterol	300 mg
Sodium	2400 mg
Carbohydrate	375 g
Protein	65 g

(U.S. Food and Drug Administration and International Food Information Council)



So, how does fast food fit into a healthy diet? Examine the information below about the nutritional content of fast-food sandwiches. With your group, study one of the nutritional components (total calories, total fat, cholesterol, sodium, carbohydrates, or protein). Use box plots, histograms, and the measures of central tendency and spread to compare the amount of that component in the sandwiches. You may even want to divide your data so that you can make comparisons between types of sandwiches (burger, chicken, or fish) or between restaurants. As you do your statistical analysis, discuss how these fast-food items would affect a healthy diet. Prepare a short report or class presentation discussing your conclusions.

Consumer CONNECTION

Fast food is a popular choice today because it is quick and convenient. Despite being high in fat, calories, sodium, and cholesterol, fast food is not bad, nutritionists say, but should be consumed in moderation with an otherwise healthy diet. Many fast-food restaurants have responded to America's new health consciousness and now offer low-calorie menu items such as salads, lean meats, and chili.

Fast Food Nutrition Facts

Sandwich	Total calories	Total Fat (g)	Cholesterol (mg)	Sodium (mg)	Carbohydrate (g)	Protein (g)
Burger King "Whopper Jr."	400	24	55	530	28	19
Carl's Jr. "Jr. Hamburger"	330	13	45	480	34	18
Dairy Queen "Hamburger"	310	13	45	580	29	17
Hardee's "Hamburger"	270	11	35	550	29	13
Jack in the Box "Hamburger"	280	12	30	490	30	12
McDonald's "Hamburger"	270	8	30	600	35	13
Wendy's "Single Hamburger"	420	20	70	920	37	25
Whataburger "Whataburger Jr."	322	13	42	603	35	16
Burger King "BK Broiler"	530	26	105	1060	45	29
Carl's Jr. "Barbecue Chicken"	280	3	60	830	37	25
Dairy Queen "Grilled Chicken Fillet"	300	8	50	800	33	25
Hardee's "Chicken Fillet"	480	23	55	1190	44	24
Jack in the Box "Chicken Sandwich"	420	23	40	950	39	16
McDonald's "Crispy Chicken"	550	27	54	1180	54	23
Wendy's "Grilled Chicken"	310	8	65	790	35	27
Whataburger "Grilled Chicken"	442	14	48	1103	66	34
Burger King "BK Big Fish"	720	43	80	1180	59	23
Carl's Jr. "Carl's Catch"	510	27	80	1030	50	18
Dairy Queen "Fish Fillet"	370	16	45	630	39	16
Hardee's "Fisherman's Fillet"	530	28	75	1280	45	25
McDonald's "Filet-O-Fish"	470	26	50	890	45	15
Whataburger "Whatacatch"	467	25	33	636	43	18

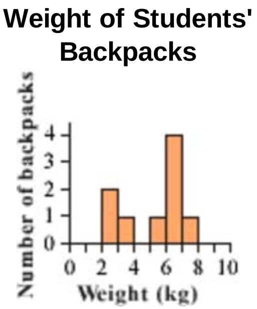
(*The NutriBase Complete Book of Food Counts*, 2001)

In order to provide a complete statistical analysis of a data set, statisticians often need to use several different measurements and graphs. Throughout this course you will learn about more statistics that help you make accurate predictions and conclusions from data.

EXERCISES

Practice Your Skills

1. The histogram at right shows a set of data of backpack weights.


Weight of Students' Backpacks

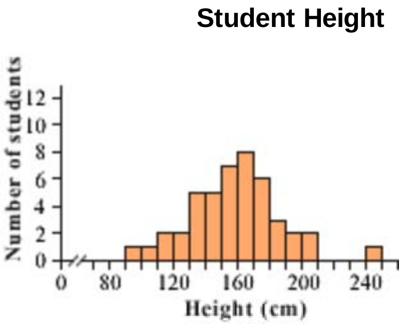
a. How many values are between 2 kg and 3 kg?

b. How many values are in the data set?

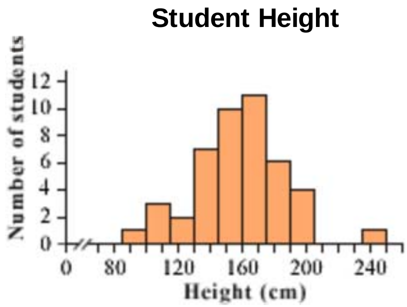
c. Make up a set of data measured to the nearest tenth of a kilogram that creates this histogram.

2. Study these four histograms.

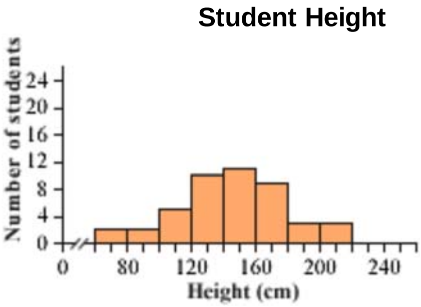
i.


Student Height

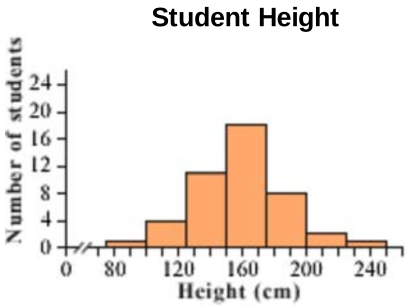
ii.


Student Height

iii.


Student Height

iv.


Student Height

a. What is the bin width of each histogram above?

b. Which histogram could not come from the same data set as the other three? Explain why.

3. These data are the head circumferences in centimeters of 20 newborn girls:

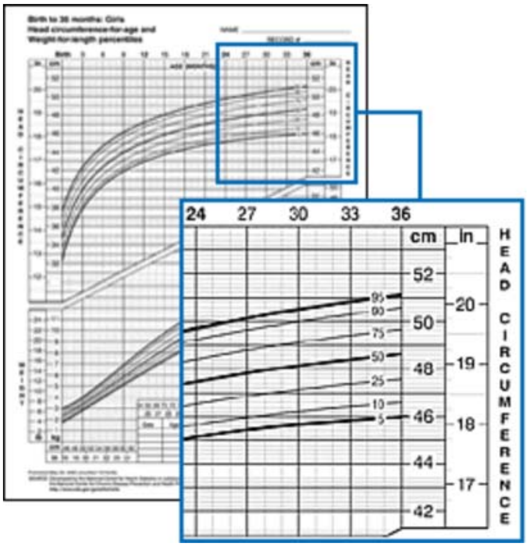
{31, 32, 33, 33, 33, 34, 34, 34, 34, 34,

35, 35, 35, 35, 35, 36, 36, 36, 37, 38}

a. How many values are below 34 cm?

b. What is the percentile rank of 34 cm?

c. What is the percentile rank of 38 cm?



Growth charts, such as the one shown at right, often give a range of data divided into percentiles. This chart shows the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles.

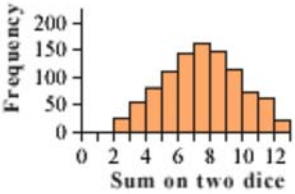
Keymath.com More Practice Your Skills

Reason and Apply

4. Carl and Bethany roll a pair of dice 1000 times and keep track of the sum on the two dice. The frequency of each sum is listed below and shown in the histogram.

Sum	Frequency
2	26
3	56
4	83
5	110
6	145
7	162
8	149
9	114
10	73
11	61
12	21

Probability Experiment

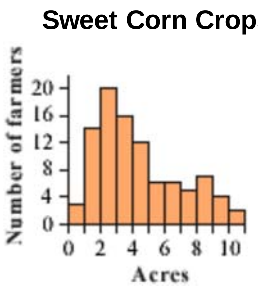


a. Graph this histogram on your calculator. List the window values needed for it to look like the histogram above.

b. Explain why the histogram is mound-shaped.

c. Describe how to find the mean sum and the median sum for this data set.

5. Rita and Noah survey 95 farmers in their county to see how many acres of sweet corn each farmer has planted. They summarize their results in a histogram.



A harvesting vehicle at work in a corn field.

a. The distribution is skewed right. Explain what this means in terms of the data set.

b. Graph the histogram on your calculator.

c. Describe what you think a box plot of this information would look like, and then check your conjecture with your calculator.

6. Describe a situation and sketch a histogram to reflect each condition named below.

a. mound-shaped and symmetric

b. skewed left

c. skewed right

d. rectangular

7. Ignacio kept a log of the amount of time he spent doing homework and watching television during 20 school days.

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Homework (min)	4	10	40	11	55	46	46	23	57	28	65	58	52	38	38	39	45	27	41	44
Television (min)	78	30	15	72	25	30	90	40	35	56	12	5	95	27	38	50	10	42	60	34

a. Draw two box plots, one showing the amount of time spent doing homework, and one showing the amount of time watching television. Which distribution has the greater spread?

b. Make an educated guess about the shape of a histogram for each set of data. Will either be skewed? Mound-shaped? Check your guess by drawing each histogram.

c. Calculate the median and interquartile range, and the mean and standard deviation, for both homework and television. Which measure of spread best represents the data?

8. At a large university, 1500 students took a final exam in chemistry.

a. Frank learns that his score of 76 (out of 100) places him at the 88th percentile. How many students scored lower than Frank? How many scored higher?

b. Mary scored 82, which placed her at the 95th percentile. Describe how Mary's performance compares to that of others in the class.

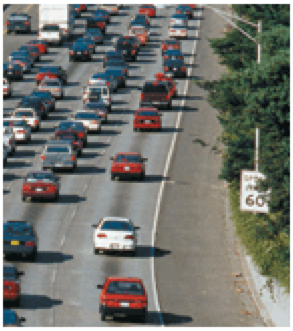
c. The highest score on the exam was 91. What percentile rank is associated with this score?

d. Every student who scored above the 90th percentile received an A. How many students earned this grade?

e. Explain the difference between a percent score and a percentile rank. In your opinion, should you be evaluated based on percent or percentile? Why?

Science CONNECTION

How are speed limits determined? Radar checks are performed at selected locations on the roadway to collect data about drivers' speeds under ideal driving conditions. A statistical analysis is then done to determine the 85th percentile speed. Studies suggest that posting limits at the 85th percentile minimizes accidents and traffic jams and that drivers are more likely to comply with the speed limit. You can learn more about how speed limits are determined by using the links at www.keymath.com/DAA.



9. **APPLICATION** Traffic studies have shown that the best speed limit to post on a given road is the 85th percentile speed.

Assume a road engineer collects these data to determine the speed limit on a local street.

Speed (mi/h)	Number of Cars (frequency)	Speed (mi/h)	Number of Cars (frequency)
13–15	1	31–33	17
16–18	2	34–36	13
19–21	5	37–39	7
22–24	11	40–42	6
25–27	15	43–45	1
28–30	21	46–48	1

(*Iowa Traffic Control Devices and Pavement Markings: A Manual for Cities and Counties,* Center of Transportation Research and Education, 2001)

a. Draw a histogram for these data.

b. Find the 85th percentile speed.

c. What speed limit would you recommend based on this traffic study?

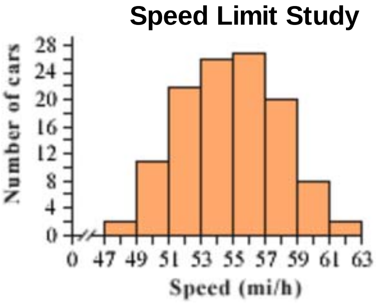
d. What other factors should be considered in determining a speed limit?

10. **APPLICATION** A road engineer studies a rural two-lane highway and presents this histogram to the County Department of Highways.

a. For how many cars was data collected?

b. What is the 85th percentile speed?

c. What speed limit would you recommend for this highway?



Review

11. Penny calculates that the deviations from the mean for a data set of eight values are 0, – 40, –78, –71, 33, 36, 42, and 91.

a. How do you know that at least one of the deviations is incorrect?

b. If it turns out that 33 is the only incorrect deviation, what should the correct deviation be?

c. Use the corrected deviations to find the actual data values, the standard deviation, the median, and the interquartile range if the mean is

i. 747 **ii.** 850

d. Write your observations from 11a–c.

12. At Piccolo Pizza Parlor, a large cheese pizza sells for $8.99. Each topping costs an additional $0.50.

a. How much will a four-topping pizza cost?

b. The Piccolo Extra Special has eight toppings and costs $12.47. How much do you save by ordering this special combination instead of ordering eight toppings of your own choosing?

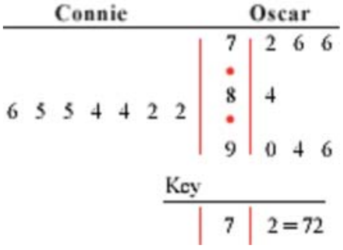
13. Courtney can run 100 m in 12.3 s. Marissa can run 100 yd in 11.2 s. Who runs faster? (There are 2.54 centimeters per inch, 12 inches per foot, and 3 feet per yard.)

Project

STEM-AND-LEAF PLOTS

John Tukey introduced **stem-and-leaf plots** in 1972 as a form for one-variable data that is appropriately compact and easy to look over.

Something like a sideways histogram, the stem-and-leaf plot is more detailed because individual data values can be found in the graph. Here are Connie's and Oscar's scores from Exercise 6 in Lesson 2.1 displayed in a stem-and-leaf plot.



Connie		Oscar
	7	2 6 6
6 5 5 4 4 2 2	8	4
	9	0 4 6

Key

	7	2 = 72

How do you think a stem-and-leaf plot works and why do you need a key? Read about stem-and-leaf plots in a high school or college statistics book and prepare a research project.

Your project should include

- Instructions to make a stem-and-leaf plot from data.
- Sample plots using data from this chapter. Include variations of stem-and-leaf plots.
- How to decide what level of accuracy is needed to communicate the data usefully.
- Insights obtained from your stem-and-leaf plots.

EXPLORATION

Census Microdata

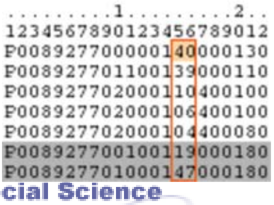
Since the establishment of the United States, the Constitution has required a nationwide population count, or census. The first U.S. census was conducted in 1790, less than a year after George Washington was inaugurated. A full census has been conducted every 10 years since.

The first censuses were primarily concerned with the number of people so that the federal government could make decisions about representation and taxation. Today, however, the U.S. Census Bureau collects a wide variety of data, including age, sex, race, national origin, marital status, and education. You can learn more about the history of the U.S. Census by visiting the Internet links at www.keymath.com/DAA .

A census taker gathers data at a New York City home in 1930.

The most detailed information published by the U.S. Census Bureau is called microdata-data about individuals. This microdata is originally published as an array of numbers, as shown below. You can see that microdata, by itself, would not be very useful to someone trying to make decisions.

Collecting data is only the first part of the U.S. Census Bureau's job. The Bureau also analyzes these data and publishes reports that help federal, state, and local governments, organizations, businesses, and citizens make decisions. In this exploration you'll use Fathom^{TM} Dynamic Statistics^{TM} software to analyze a set of census microdata. You'll make some conjectures about the data set and use what you've learned about statistics in this chapter to support or refute your conjectures.



```
.........1.........2..
1234567890123456789012
P008927700000140000130
P008927701100139000110
P008927702000110400100
P008927702000106400100
P008927702000104400080
P008927700100119000180
P008927701000147000180
```

This table shows 1990 U.S. Census microdata about people living around Berkeley, California. Each row represents one individual. Shaded and unshaded rows group individuals that live in the same household. Columns represent specific information about each individual. For example, columns 15-16 indicate age. In the first row the person is 40 years old.

Social Science CONNECTION

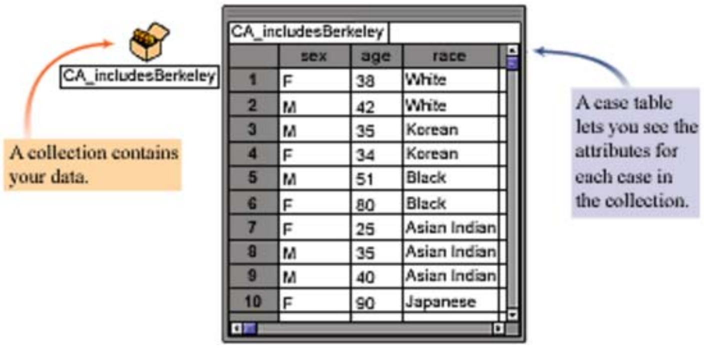
In theory, the census counts every person living in the United States. In actuality, despite outreach programs that attempt to count everyone, including people without housing and people who are not able to read or complete the census, the 2000 U.S. Census is estimated to have excluded up to 3.4 million people. Furthermore, the U.S. Census Bureau today uses both a short form and a long form, such that some questions are only asked to a small sample (5% to 20%) of the population, and the results are statistically applied to the whole population. Some statisticians believe that a census founded entirely on random sampling, but conducted more thoroughly by tracking down every single person in that sample, may be more beneficial in the future.

Activity

Different Ways to Analyze Data

Step 1 Start Fathom. From the File menu choose **Open.** Open one of the census data files in the **Sample Documents** folder. You'll see a box of gold balls, called a collection, that holds data about several individuals, or cases.

Step 2 Click on the collection and then choose **Case Table** from the Insert menu. You now have a table of your microdata. Scroll through the data. How many people are there? What specific data, or attributes, were collected about each person?



CA_includesBerkeley

A collection contains your data.

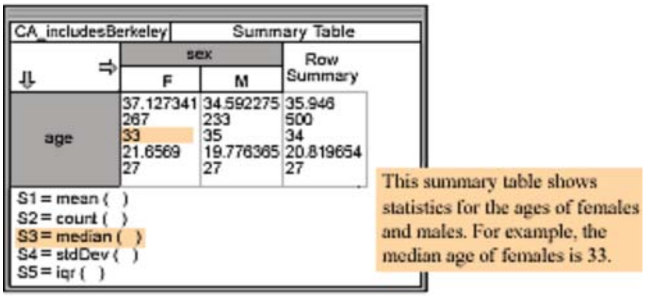
CA_includesBerkeley

	sex	age	race
1	F	38	White
2	M	42	White
3	M	35	Korean
4	F	34	Korean
5	M	51	Black
6	F	80	Black
7	F	25	Asian Indian
8	M	35	Asian Indian
9	M	40	Asian Indian
10	F	90	Japanese

A case table lets you see the attributes for each case in the collection.

Step 3 Make a conjecture about the people in your data set. Your conjecture can be about just one attribute, such as "The majority of people have some education beyond high school," or it can be about a combination of attributes, such as "The males are older than the females." Your conjecture should be something that you can test and, therefore, support or refute with statistics.

Step 4 Begin testing your conjecture by calculating summary statistics, such as the mean, median, standard deviation, and interquartile range. Choose **Summary Table** from the Insert menu, and then drag and drop attributes from your case table. From the Summary menu choose **Add Basic Statistics** to see some of the statistics that you have studied in this chapter. Based on these statistics, does your conjecture seem to be true? Why or why not?



CA_includesBerkeley Summary Table

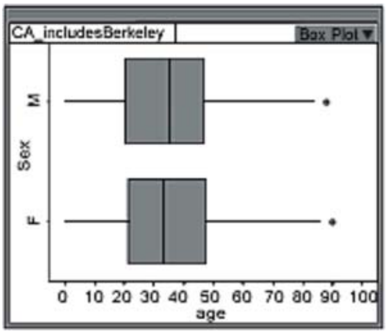
⇩ ⇨	sex		Row Summary
	F	M	
age	37.127341 267 33 21.6569 27	34.592275 233 35 19.776365 27	35.946 500 34 20.819654 27

S1 = mean ( )
S2 = count ( )
S3 = median ( )
S4 = stdDev ( )
S5 = iqr ( )

This summary table shows statistics for the ages of females and males. For example, the median age of females is 33.

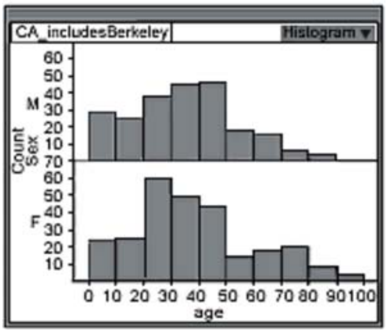
Step 5

Graphs are another way to test your conjecture. Choose **Graph** from the Insert menu. Drag and drop attributes into your graph and then choose **Box Plot** from the pull-down menu in the corner of the graph window. Do box plots help you support or refute your conjecture?



Step 6

Now create histograms to test your conjecture. You can either follow the process in Step 5 to create a new graph or you can simply use the pull-down menu to change your box plots to histograms. What new information do the histograms give you? Do histograms help support your conjecture? Do the histograms better support your conjecture if you change the bin width?



Step 7

Look at all of your analyses, including the summary statistics and graphs. Do you think your conjecture is true?
If you aren't sure, you might want to modify your conjecture, or think about factors that might also be at work. Write a short paragraph explaining your findings. Think about these questions as you write: What factors might affect your analysis? Can you explain any outliers in your data? Which statistic or graph revealed the most about these data? In what ways might citizens or governments use these data to make informed decisions?

Questions

1. In this exploration, you've seen some ways to determine whether or not a conjecture is true. Is it possible for a conjecture to appear to be true or false, depending on what statistic or graph you select? Make a new conjecture for your microdata and try to find one statistic or graph that supports the conjecture and one that refutes the conjecture.

2. State and federal decision-makers often have to compare data from different regions to make sure they are meeting everyone's needs. Use Fathom to compare microdata for two different geographic regions. Make and test a few conjectures about how the regions compare and contrast. Describe at least one way in which these communities could use your graphs to make decisions.

CHAPTER 2 REVIEW

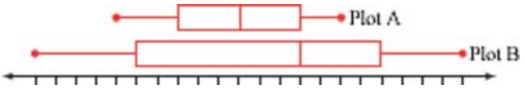
Sets of data, such as temperatures, the sugar content of breakfast cereals, and the number of hours of television you watch daily, can be analyzed and pictured using tools that you learned about in this chapter. The **mean** and the **median** are **measures of central tendency.** They tell you about a typical value for the data set. But a measure of central tendency alone does not tell the whole story. You also need to look at the **spread** of the data values. The **standard deviation** helps you determine spread about the mean, and the **interquartile range** helps you determine spread about the median. These measures of spread are also frequently used to identify **outliers** in the data set.

To display a data set visually, you can use a box plot or a histogram. A **box plot** shows the median of the data set, the **range** of the entire set, and the interquartile range between the **first quartile** and the **third quartile.** A **histogram** uses **bins** to show how the data are spread throughout the entire range. By changing the width of the bins, you can get a different perspective on the distribution. Neither a box plot nor a histogram shows individual data values, but both help you see whether the data set is **symmetric** or **skewed.**

Percentile ranks are useful to show how one data value compares to the data set as a whole. The percentile rank of one data value tells you the percentage of values that are less than the given data value. Percentile ranks are not the same as percent scores.

By using a combination of **statistics** and graphs, you can better understand the meaning and implications of a data set. A careful analysis of a data set helps you make general conclusions about the past and predictions for the future.

EXERCISES



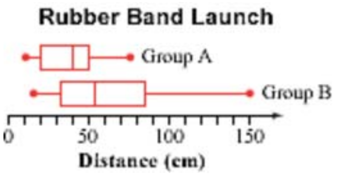
1. Which box plot has the greater standard deviation? Explain your reasoning.

2. Consider these box plots. Group A conducted the rubber band launch experiment 30 times, and Group B conducted the experiment 25 times.

a. How many data values are represented in each whisker of each box plot?

b. Which data set has the greater standard deviation? Explain how you know.

c. Draw two histograms that might represent the information pictured in these two box plots.



3. The Los Angeles Lakers won the 2002 National Basketball Association Championship. This table gives the total points scored by each player during that season.

Total Points Scored by Los Angeles Lakers Players (2001-2002 Season)

Player	Points	Player	Points	Player	Points
Kobe Bryant	2019	Robert Horry	550	Shaquille O'Neal	1822
Joseph Crispin	10	Lindsey Hunter	473	Mike Penberthy	5
Derek Fisher	786	Mark Madsen	167	Mitch Richmond	260
Rick Fox	645	Jelani McCoy	26	Brian Shaw	169
Devean George	581	Stanislav Medvedenko	331	Samaki Walker	460

(www.nba.com)

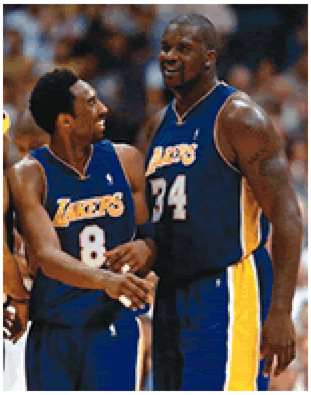
a. Find the mean, median, and mode for this data set.

b. Find the five-number summary.

c. Draw a box plot of the data. Describe the shape of the data set.

d. Calculate the interquartile range.

e. Identify any outliers. Use the 1.5 · *IQR* definition for outliers.


Kobe Bryant and Shaquille O'Neal

4. Invent two data sets, each with seven values, such that Set A has the greater standard deviation and Set B has the greater interquartile range.

5. The table below contains the recorded extreme temperatures, in degrees Fahrenheit, for each of the seven continents.

Highest and Lowest Recorded Temperatures

Continent	High (°F)	Low (°F)
Africa	136	–11
Antarctica	59	–129
Asia	129	–90
Australia	128	–9
Europe	122	–67
North America	134	–87
South America	120	–27

(www.infoplease.com)

a. Find the mean and standard deviation of the high temperatures.

b. Find the mean and standard deviation of the low temperatures.

c. Which temperatures, if any, are outliers in each set of data? Use the 2*s* definition for outliers.

6. Below is a list of Academy Award winners in the Best Actress and Best Actor categories and each person's age when he or she received the award.

Academy Award Winners 1970-2001

Year	Best actress in a leading role, Age	Best actor in a leading role, Age	Year	Best actress in a leading role, Age	Best actor in a leading role, Age
1970	Glenda Jackson, 34	George C. Scott, 43	1986	Marlee Matlin, 21	Paul Newman, 62
1971	Jane Fonda, 34	Gene Hackman, 42	1987	Cher, 41	Michael Douglas, 43
1972	Liza Minnelli, 26	Marlon Brando, 48	1988	Jodie Foster, 26	Dustin Hoffman, 51
1973	Glenda Jackson, 37	Jack Lemmon, 49	1989	Jessica Tandy, 81	Daniel Day-Lewis, 32
1974	Ellen Burstyn, 42	Art Carney, 56	1990	Kathy Bates, 42	Jeremy Irons, 42
1975	Louise Fletcher, 41	Jack Nicholson, 38	1991	Jodie Foster, 29	Anthony Hopkins, 54
1976	Faye Dunaway, 36	Peter Finch, 60	1992	Emma Thompson, 33	Al Pacino, 52
1977	Diane Keaton, 32	Richard Dreyfuss, 30	1993	Holly Hunter, 36	Tom Hanks, 37
1978	Jane Fonda, 41	Jon Voight, 40	1994	Jessica Lange, 45	Tom Hanks, 38
1979	Sally Field, 33	Dustin Hoffman, 42	1995	Susan Sarandon, 49	Nicolas Cage, 31
1980	Sissy Spacek, 31	Robert De Niro, 37	1996	Frances McDormand, 39	Geoffrey Rush, 45
1981	Katharine Hepburn, 74	Henry Fonda, 76	1997	Helen Hunt, 34	Jack Nicholson, 60
1982	Meryl Streep, 33	Ben Kingsley, 39	1998	Gwyneth Paltrow, 26	Roberto Benigni, 46
1983	Shirley MacLaine, 49	Robert Duvall, 53	1999	Hilary Swank, 25	Kevin Spacey, 40
1984	Sally Field, 38	F. Murray Abraham, 45	2000	Julia Roberts, 33	Russell Crowe, 36
1985	Geraldine Page, 61	William Hurt, 36	2001	Halle Berry, 33	Denzel Washington, 47

(*www.imdb.com*)

a. Find the mean age and the median age for the Best Actress winners.

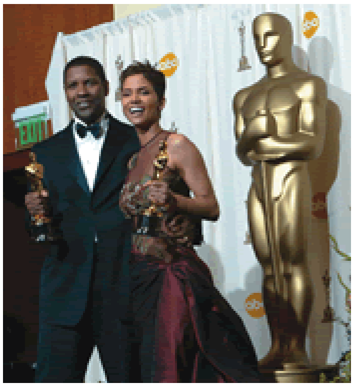
b. Find the mean age and the median age for the Best Actor winners.

c. On the same axis, draw two box plots, one for the age of Best Actress winners and the other for the age of Best Actor winners.

d. Draw two histograms, one each for Best Actress and Best Actor.

e. Use your graphs from 6c and 6d to predict which data set has the greater standard deviation. Explain your reasoning. Then calculate the standard deviations to check your prediction.

f. Julia Roberts was 33 years old when she won Best Actress in 2000. What is her percentile rank among all Best Actress winners from 1970 to 2001? Explain what this percentile rank tells you.



At the 74th annual Academy Awards, Halle Berry was the first African-American to win Best Actress. Denzel Washington was the second African-American to win Best Actor.

7. **APPLICATION** Following the 1998 Academy Awards ceremony, Best Actress nominee Fernanda Montenegro (age 70) said Gwyneth Paltrow (age 26) had won Best Actress for *Shakespeare in Love* because she was younger. This comment caused Pace University student Michael Gilberg and professor Terence Hines to test the theory that younger women and older men are more likely than older women and younger men to receive an Academy Award for Best Actress and Best Actor. Their study was published in the February 2000 issue of *Psychological Reports*.

Assume that you are working with Gilberg and Hines. Use your statistics and graphs from Exercise 6 to confirm or refute the theory that younger women and older men are more likely to win. Prepare a brief report on your conclusions.

8. The 2000 U.S. passenger-car production totals are shown at right.

a. Make a box plot of these data.

b. Make a histogram using a bin width that provides meaningful information about the data.

c. Suppose a different year has a similar distribution but the total number of cars produced is 400 thousand greater than in 2000. Describe how this could affect the shape of your box plot and histogram.

d. What is the percentile rank of Pontiac?

e. What is the percentile rank of Ford?

2000 U.S. Passenger-Car Production

Brand	Number of cars (thousands)
BMW	39
Buick	209
Cadillac	159
Chevrolet	597
Chrysler	80
Dodge	298
Ford	965
Honda	677
Lincoln/Mercury	282
Mazda	107
Mitsubishi	222
Nissan	150
Oldsmobile	235
Plymouth	55
Pontiac	576
Saturn	261
Subaru	108
Toyota	520

(The World Almanac and Book of Facts 2003)

TAKE ANOTHER LOOK

1. Which measure of central tendency do you think of when someone says "average"? Without clarification, an "average" could be the mean, the median, or the mode. Find newspaper and magazine articles that state an "average," such as "The average American family has 2.58 children." Read the articles closely and analyze any data that are provided. Can you find enough information to tell which measure of central tendency is being used? Do you find that most articles are or are not specific enough with their mathematics? What conclusions can you make?

2. The calculation of mean that you learned in this chapter-the sum of all data values divided by the number of values-is more precisely called the **arithmetic mean.**

Other means include the geometric mean, the harmonic mean, the quadratic mean, the trimean, and the midmean. Research one or more of these means and compare and contrast the calculation to the arithmetic mean.

3. In Lesson 2.3, you learned how to find the median of a data set by looking at a histogram. How would you use the histogram to approximate the mean? The mode?

4. Another measure of spread, the mean deviation, *MD*, uses absolute value to eliminate the effect of the different signs of the individual deviations.

$$MD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Try using mean deviation for some of the exercises in which you calculated standard deviation. How do the values compare? When might standard deviation or mean deviation be the more appropriate measure of spread?

Assessing What You've Learned

BEGIN A PORTFOLIO An artist usually keeps both a notebook and a portfolio. The notebook might contain everything from scratch work to practice sketches to notes about past or future subject matter. The portfolio, in contrast, is reserved for the artist's most significant or best work.

As a student, you probably already keep a notebook that contains everything from your class notes to homework to research for independent projects. You can also start a separate portfolio that collects your most significant work.

Review all the work you've done so far and find your best works of art: the neatest graphs, the most thorough calculations for various statistics, the most complete analysis of a data set, or the most comprehensive project. Add each piece to your portfolio with a paragraph or two that addresses these questions:

- What is the piece an example of?
- Does this piece represent your best work? Why else did you choose it?
- What mathematics did you learn or apply in this piece?
- How would you improve the piece if you redid or revised it?

WRITE TEST ITEMS Writing your own problems is an excellent way to assess and review what you've learned. If you were writing a test for this chapter, what would it include? Start by having a group discussion to identify the key ideas of the chapter. Then divide the lessons among group members, and have each group member write at least one problem for each lesson assigned to him or her. Try to create a mix of problems, from simple one-step exercises that require you to recall facts and formulas, to complex multistep problems that require more thinking. Because you'll be working with data and statistics, you'll need to carefully consider which statistics and graphs are appropriate for the data.

Share your problems with your group members and try out one another's problems. Then discus the problems in your group:

- Were the problems representative of the content of the chapter?
- Were any problems too hard or too easy?
- Were the statistics appropriate for the data?